SIN-CHN SCIENTIFIC PRESS PTE.LTD.

Molecular & Cellular Biomechanics 2025, 22(3), 1150.
https://doi.org/10.62617/mcb1150

Article

Application of machine classification learning models based on factor space mathematical theory in higher vocational education from a biomechanical perspective

Wei Luo[1], Guo-qing Liu[2], Zi-jian Wu[1], Zhi-yan Sa[1,*]

[1] A Ba Vocational College, Aba Tibetan and Qiang Autonomous Prefecture 623200, China
[2] Malkang Education Bureau, Aba Tibetan and Qiang Autonomous Prefecture 624000, China
*** Corresponding author:** Zhi-yan Sa, abzyszy1@163.com

CITATION

Luo W, Liu G, Wu Z, Sa Z. Application of machine classification learning models based on factor space mathematical theory in higher vocational education from a biomechanical perspective. Molecular & Cellular Biomechanics. 2025; 22(3): 1150. https://doi.org/10.62617/mcb1150

ARTICLE INFO

Received: 18 December 2024
Accepted: 9 January 2025
Available online: 18 February 2025

Abstract: With the rapid advancement of biomechanical research and educational big data, there is a growing need to integrate sophisticated analytical tools to enhance the understanding of human movement, learning behaviors, and their interactions. Traditional machine learning models often fall short in capturing the complex, multi-dimensional relationships inherent in biomechanical and educational datasets, leading to limited precision, inadequate personalization, and poor generalization capabilities, which restrict their applicability in dynamic teaching environments. To address these challenges, this paper proposes a machine learning model based on Factor Space Mathematical Theory integrated with the extreme gradient boosting (XGBoost) algorithm. By leveraging Factor Space Mathematical Theory, the model effectively captures the multi-dimensional characteristics of biomechanical and educational data, addressing the oversimplification and unidimensional nature of traditional models. Moreover, with the robust classification and prediction performance of XGBoost, the proposed model enhances the ability to generalize and process complex educational data. Experimental results demonstrate that the proposed model achieves an accuracy of 0.92 and an *F*1 score of 0.90 in predicting students' biomechanical performance metrics, such as gait analysis and posture stability, which are critical for understanding learning behaviors in physical education and vocational training. The model outperforms the standalone XGBoost model by a significant margin of 0.05 in accuracy. Additionally, MSE analysis across diverse datasets reveals no evidence of overfitting, further validating the model's strong generalization capabilities. This study highlights the effectiveness of combining Factor Space Theory with XGBoost, offering improved accuracy, operational efficiency, and adaptability in biomechanical data analysis and educational behavior prediction. The findings provide a novel perspective and practical approach to advancing biomechanical research and its application in educational reform, particularly in higher vocational education.

Keywords: educational big data; machine learning; factor space mathematical theory; extreme gradient boosting; higher vocational education reform

1. Introduction

As an essential part of vocational education, higher vocational education bears the responsibility of cultivating a large number of skilled and application-oriented talents to meet the demands of a dynamic society [1]. With the ongoing transformation and upgrading of the economic structure, as well as the increasing demand for highly skilled professionals, higher vocational colleges have taken on an increasingly vital role in enhancing students' professional quality and employability. In this context, the adoption of advanced technologies, such as big data [2], artificial intelligence [3], and

machine learning [4–6], has opened up new avenues for addressing challenges in teaching quality and management precision. Leveraging these technologies for educational reform has become an important topic in current research.

Machine learning, as a robust data analysis and pattern recognition tool, has achieved remarkable outcomes across many industries, and its potential in education is gradually being realized. Specifically, in areas such as student performance prediction, learning behavior analysis, and personalized learning recommendations, machine learning has demonstrated significant promise. For instance, Berens et al. [7] developed an early detection system for identifying students at risk of dropout by employing the AdaBoost algorithm along with regression analysis, neural networks, and decision trees. Their approach improved prediction accuracy from 79% to 90% for public universities and from 85% to 95% for private universities over four semesters. Similarly, Ikawati et al. [8] proposed a learning style prediction model using an ensemble tree method that integrates bagging and gradient-boosted trees, achieving higher classification accuracy compared to single tree models. Ouatik et al. [9] applied KNN, C4.5, and SVM algorithms to predict student success, with the SVM algorithm attaining a prediction accuracy of 87.32%. Wu et al. [10] used artificial neural networks (ANNs) and support vector machines (SVMs) to diagnose students with learning disabilities, demonstrating that ANNs outperformed SVMs in recognition accuracy.

Moreover, personalized learning has also gained traction. Amin et al. [11] developed a personalized e-learning and MOOC recommender system using an intelligent electronic platform, which collects data on students' performance, interests, and learning preferences to recommend suitable courses. While these studies highlight the potential of machine learning in educational applications, most existing models focus on algorithm universality and prediction accuracy, neglecting the comprehensive modeling of multidimensional factors in education. For higher vocational education, factors such as students' personal characteristics, learning behaviors, teachers' teaching methods, and curriculum design require integrated and nuanced modeling approaches.

Recent research has attempted to address these gaps by incorporating diverse theories and methodologies into educational data analysis. For example, Alshurafat et al. [12] explored the impact of online learning systems on accounting students using an integrated model based on social capital theory, rational behavior theory, and the technology acceptance model, identifying social trust as a key factor influencing perceived usefulness and ease of use. Delen et al. [13] developed a Bayesian belief network-based model to predict student attrition, highlighting conditional dependencies and interrelationships among factors. Li et al. [14] proposed a comprehensive evaluation mechanism for physical education teaching quality using multivariate data and achieved an accuracy of over 97%. Godwin et al. [15] introduced topological data analysis as a people-oriented approach to address quantitative methods in engineering education, while Mubarak et al. [16] modeled students' performance using graph convolutional networks to capture semantic relationships in massive online learning data. Similarly, Yakubu and Dasuki [17] adopted the Unified Theory of Technology Acceptance to identify convenience and behavioral intention as significant factors influencing e-learning adoption.

Despite these advancements, challenges persist in applying complex models to higher vocational education. Specifically, integrating factor space mathematical theory into machine learning for analyzing large-scale educational data presents difficulties in balancing computational efficiency with prediction accuracy. Factor space mathematical theory offers a structured approach to model the multidimensional and highly correlated nature of educational data, while XGBoost [18–20], a powerful machine learning algorithm, excels at processing large datasets with robust classification and prediction capabilities. For instance, Osman et al. [18] demonstrated the efficacy of XGBoost in groundwater level prediction, and Kavzoglu et al. [19] highlighted its superior performance in landslide susceptibility mapping. Furthermore, applications of SVMs [21,22] and decision tree models [23,24] in biomechanics and education underscore the potential of integrating domain-specific knowledge into predictive analytics.

This study combines factor space mathematical theory with the XGBoost algorithm to develop a novel model for higher vocational education. By extracting potential influencing factors through mathematical modeling and leveraging XGBoost's ability to process massive student data, this approach provides accurate predictions and classifications to support student evaluation and teaching management. Experimental comparisons with traditional models, including standalone XGBoost, SVM, and decision trees, show that the proposed model outperforms these alternatives in accuracy, precision, recall, and $F1$ score. Specifically, the proposed model achieves an accuracy of 0.92 and an $F1$ score of 0.90, demonstrating superior generalization ability and avoiding overfitting across different datasets.

The findings suggest that this integrated approach not only identifies the relationship between student characteristics and grades but also offers practical insights for teaching reforms in higher vocational colleges. The model's predictions can help teachers monitor students' learning progress and provide valuable data support for institutional decision-making, paving the way for data-driven educational transformations.

2. Modeling the XGBoost model by introducing the factor space theory

2.1. Multidimensional factor space modeling

In this study, the factor space theory is introduced to model multidimensional factors, aiming to integrate data from various dimensions into a high-dimensional space. This approach enables the capture of complex relationships within the data. In the context of higher vocational education, multidimensional factors encompass a wide range of data, including students' personal background information, learning behavior records, teachers' teaching characteristics, and course-specific attributes. Each of these factors can be regarded as an independent feature; however, these features rarely operate in isolation, as intricate interactions often exist between them.

To address this, the study employs feature engineering to construct and model these interactions effectively. By leveraging factor space theory, the proposed

approach facilitates the identification and representation of latent relationships among these features, thus enabling a more comprehensive analysis of the underlying educational data. This method ensures that the intricate interdependencies between various dimensions, such as the impact of teaching strategies on students with different backgrounds or the influence of course design on learning behaviors, are adequately captured and utilized in the modeling process. Consequently, the integration of factor space theory offers a robust framework for tackling the complexity inherent in educational data, laying the foundation for more accurate and insightful analysis in higher vocational education.

Table 1 is a partial display of the feature categories that may exist in higher vocational education and the feature names and feature data types corresponding to the features. The table shows student background features, learning behavior features, teacher features, and interactive features. Different feature data types may be different. For example, the age of students and the length of study are numerical types and can be directly input into the model for processing. However, some features, such as student gender and course difficulty, are categorical data types and cannot be directly input into the model for unified processing.

Table 1. Characteristics table.

Feature Category	Feature Name	Feature Type
Student Background Features	Student Gender	Categorical variables
	Student Age	Continuous variables
Learning behavior Features	Online learning duration	Continuous variables
	Class participation	Continuous variables
Teacher Features	Teaching Method	Categorical variables
	Teacher Evaluation	Categorical variables
Interaction Features	Study time and participation	Continuous variables

This paper processes categorical data by converting them into numerical types through one-hot encoding [25]. The specific approach is to convert each category value into a binary bit to form a vector representation of the numerical value. For example, for the gender feature of students, this feature has two different categories (male and female). For each category, it is converted into a vector representation of the numerical value:

$$v_{\text{male}} = [1,0] \tag{1}$$

$$v_{\text{female}} = [0,1] \tag{2}$$

In this way, different categories of different features can be represented by unique numerical representations. Then the input feature matrix can be constructed based on the data. Assuming there are m students and n features, the size of the input feature matrix X is $m \times n$. The representation of the feature matrix X is:

$$\mathrm{X} = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix} \tag{3}$$

In Equation (3), x_{ij} represents the value of the *i*-th student on the *j*-th feature. At the same time, in order to capture the interaction between different features, a combination of features can be constructed for any two possibly related features. For example, for any two features x_j and x_k, their interactive features are expressed as:

$$x_{jk} = x_j \cdot x_k \tag{4}$$

In this way, a new feature x_{jk} is obtained, which is convenient for better capturing the relationship between the two features.

Since the relationship between the original features is often nonlinear, this paper studies the use of nonlinear kernel functions to map the original data of the input original features into a high-dimensional space. Through this mapping method, the nonlinear relationship of the original data in this space will become linear.

This paper adopts the Gaussian radial basis kernel [26], and calculates the inner product through the kernel function instead of explicitly calculating the mapped feature space, which can reduce the complexity of high-dimensional space calculation. Select any two original feature input data x and x', the Gaussian radial basis kernel is:

$$K(x, x') = \exp(-\frac{||x - x'||}{2\sigma^2}) \tag{5}$$

Among them, $||x - x'||$ is the Euclidean distance between the input data x and x', σ is the parameter of the kernel function. The calculation formula of the Euclidean distance is:

$$d(x, x') = \sqrt{(x_1 - x_1')^2 + (x_2 - x_2')^2 + \cdots + (x_n - x_n')^2} \tag{6}$$

Figure 1 shows the mapping of student Zhang San's data in high-dimensional space.

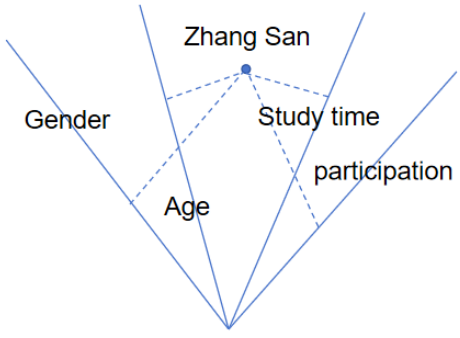

Figure 1. High-dimensional space mapping diagram.

2.2. Dimensionality reduction

When the dimensionality of data increases, it often introduces a series of challenges that can hinder the effectiveness of the model. First, high-dimensional data significantly escalates computational costs and storage requirements, which negatively impacts the efficiency of model processing and dataset training. Second, higher feature dimensions increase the risk of the model "overfitting" to the noise present in the training data, thereby reducing its generalization ability and leading to poor performance on unseen data.

To mitigate these issues and to extract the most informative components from high-dimensional data while minimizing redundancy and noise, this study employs the Principal Component Analysis (PCA) method [27,28]. PCA is a widely used dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional space by identifying the principal components that capture the majority of the data's variance. By doing so, PCA not only simplifies the data structure but also enhances model efficiency by focusing on the essential patterns in the data while discarding less significant or noisy dimensions.

In the context of this study, PCA plays a crucial role in ensuring that the model processes a compact and meaningful representation of the high-dimensional data. This step not only reduces computational overhead but also minimizes the risk of overfitting, thereby improving the robustness and accuracy of the proposed model in analyzing complex, multidimensional educational datasets.

PCA is a linear dimensionality reduction method. Its main dimensionality reduction idea is to transform the data linearly and project the data into a new coordinate system so that the new features have the largest variance. The specific steps of PCA are as follows:

1) Standardized data: First, standardize each feature degree so that the mean of the feature is 0 and the variance is 1. The input feature data matrix is X, the dimension of X is n, and the standardization formula for each feature x_i in X is:

$$x_i = \frac{x_i - \mu_i}{\sigma_i} \tag{7}$$

2) The covariance matrix [29] describes the correlation between features. The covariance matrix $\sum$ is calculated as:

$$\sum = \frac{1}{m} X^T X \tag{8}$$

In Equation (8), X is the input feature matrix, m is the number of samples, and $X^T X$ is the product of the input feature matrix and the transposed input feature matrix.

3) By performing eigenvalue decomposition on the covariance matrix $\sum$, the eigenvalues and eigenvectors can be obtained:

$$\sum v = \lambda v \tag{9}$$

In Equation (9), λ is the eigenvalue, v is the eigenvector. The eigenvalue represents the "importance" of each eigenvector. The larger the eigenvalue, the larger the variance in that direction, indicating that the feature is more important.

4) Select the k eigenvectors with the largest eigenvalues, that is, the most "important" k features, to form a matrix V_k. The dimensions corresponding to these eigenvectors will form a new low-dimensional space.

5) Finally, project the original input data onto these k principal components to obtain the reduced-dimensional data:

$$X_{\mathrm{PCA}} = XV_k \tag{10}$$

In Equation (10), V_k is the matrix composed of the first k eigenvectors, X_{PCA} is the data after dimensionality reduction.

Through the PCA method, the original input features are reduced from n dimensions to k dimensions, and most of the variance (i.e., important features) in the original input features are retained in the first k principal components, which effectively reduces the dimension of the data and removes redundant information. At the same time, the dimensionality-reduced data can be used as an effective input for the XGboost model.

2.3. XGBoost model construction

XGBoost is an integrated learning model based on the gradient boosted tree [30,31] (Gradient Boosted Trees, GBT) algorithm. In this study, the factor space modeling maps the dimensionality-reduced data into the XGBoost model through a series of decision trees to minimize the model's loss function, and introduces regularization [32] to prevent the occurrence of overfitting [33] problems.

The goal of XGBoost is to optimize the model by continuously adding trees. Its objective function mainly consists of two parts, the loss function and the regularization term. The loss function is used to measure the difference between the current model prediction value and the true value. The regularization term is to control the complexity of the model and avoid overfitting. The objective function of the XGBoost model in this article is expressed as:

$$\Gamma(F) = \sum_{i=1}^{m} l(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_k) \tag{11}$$

In Equation (11), $\Gamma(F)$ is the objective function, $l(y_i, \hat{y_i})$ is the loss function, which measures the difference between the true value y_i and the predicted value $\hat{y_i}$, $\Omega(f_k)$ is the regularization term, which is used to control the model complexity, and K is the number of trees.

The loss function studied in this paper is the mean square error (MSE):

$$l(y_i, \hat{y_i}) = \frac{1}{n} \sum_{i=1}^{n} (y_i, \hat{y_i})^2 \tag{12}$$

In Equation (12), y_i is the true value of the sample, and $\hat{y_i}$ is the predicted value of the sample.

In order to prevent overfitting, the regularization term designed in this paper mainly consists of two parts. One part is to control the complexity of the model by limiting the complexity of the number of trees (K) to prevent overfitting. On the other hand, by penalizing the size of the leaf node weight, it prevents overfitting caused by excessive weight. The regularization term of the model can be expressed as:

$$\Omega(f) = \gamma K + \frac{1}{2} \lambda \sum_{j=1}^{K} \omega_j^2 \tag{13}$$

In Equation (13), K is the total number of trees, also known as the number of iterations during model training, γ is a hyperparameter that controls the complexity of the tree, ω_j is the number of penalty trees, λ is the weight of the j-th leaf node, is the regularization coefficient, and controls the size of the leaf node weight.

The settings of XGBoost hyperparameters in this paper are shown in **Table 2**.

Table 2. Hyperparameter settings.

Hyper Parameters	Function	Value
learning_rate	Controls the contribution of each tree to the final prediction.	0.1
n_estimators	The number of weak learners.	100
max_depth	controlling the complexity of the model.	7
min_child_weight	Controls tree splitting	1
subsample	The sampling ratio of the sample	0.8
Regularization parameter	Control model complexity.	2

Table 2 sets the hyperparameters of the model, where the learning rate is set to 0.1, the number of trees is set to 100, the maximum depth of the tree is set to 6, the minimum weight of each child node is set to 1, the sample collection ratio is set to 0.8, and the regularization parameter is set to 2.

2.4. Gradient boosting algorithm

The core of the XGBoost model designed in this paper is the gradient boosting algorithm. The idea of this algorithm is to gradually reduce the value of the loss function in each iteration by adding each tree to the existing model. More specifically, the gradient boosting tree studied in this paper updates the model by gradually fitting the residual. The following are the steps for updating.

(1) For the task of the model, the model is initialized first, and the initialized data is the mean of the training data:

$$\hat{y}_i^{(0)} = \frac{1}{m}\sum_{i=1}^{m} y_i \tag{14}$$

In Equation (14), m is the number of training samples, and y_i is the true value of the *i*-th sample.

(2) In each round of iteration, XGBoosst will calculate the gradient of the current model. The gradient is equivalent to the derivative of the loss function with respect to the current predicted value. The gradient is used to indicate the error of each data point under the current model. Assuming that the current model has been iterated to round $t - 1$, and the current predicted value is $\hat{y}_i^{(t-1)}$, then in the *t*-th round of iteration, the gradient calculation is:

$$g_i^{(t)} = \frac{\partial}{\partial \hat{y}} l(y_i, \hat{y}_i^{(t-1)}) \tag{15}$$

In Equation (15), $l(y_i, \hat{y}_i^{(t-1)})$ is the loss function of the t – 1th round, and $g_i^{(t)}$ is the gradient of the *i*-th sample in the *t*-th round iteration.

(3) In addition to calculating the first-order gradient, this study also calculates the second-order gradient, which can provide information about the curvature of the loss function [34], so that the XGBoost model can perform more accurate model updates through the second-order information. The calculation formula for the second-order gradient is:

$$h_i^{(t)} = \frac{\partial^2}{\partial \hat{y}^2} l(y_i, \hat{y}_i^{(t-1)}) \tag{16}$$

(4) After calculating the gradient and second-order gradient, the XGBoost model uses a new tree to fit the current model residual Specifically, the goal of each tree in the XGBoost model is to fit the negative gradient, that is, $-g_i^{(t)}$. Therefore, this paper updates the model by minimizing the Taylor expansion of the loss function. Assuming that the current model has been iterated to round $t-1$, the current prediction value is $\hat{y}_i^{(t-1)}$, and a new tree $f_t(x_i)$ is added at this time, then the calculation formula for the new prediction value is:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i) \tag{17}$$

In Equation (17), η is the learning rate, which controls the contribution of each tree to the final model. In each iteration, the model is updated by continuously adding trees to achieve the effect of gradient improvement.

As illustrated in **Figure 2**, the operational flowchart of the gradient boosting algorithm studied in this paper is presented. The process begins with the initialization of the model. Following this, the gradient is calculated, including both the first-order and second-order gradients, which are used to measure the direction and curvature of the loss function. Based on these gradient values, the residuals are fitted, and the model is updated iteratively to minimize the residual errors.

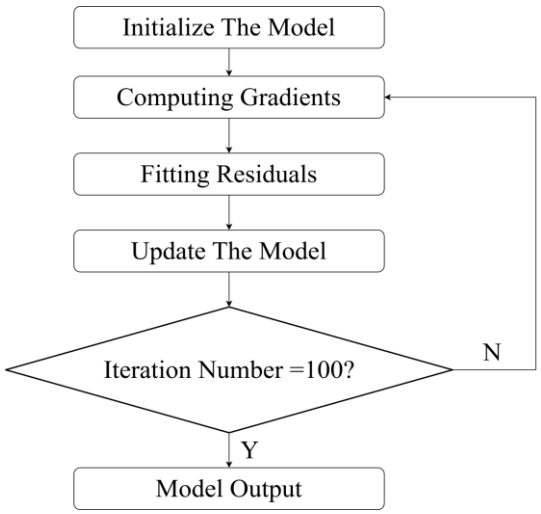

Figure 2. Gradient boosting algorithm flow chart.

Next, the algorithm checks whether the predetermined number of iterations (set to 100 in this study) has been reached. If the maximum number of iterations has been completed, the model terminates, and the final result is output. If not, the algorithm loops back to perform a new round of gradient calculations and model updates. This iterative approach ensures that the model continuously improves its predictive accuracy by reducing the residual errors at each step, ultimately achieving an optimized solution.

3. Model experimental test

3.1. Experimental data

The experiment uses the public data set Student Performance Data, which includes student information, student behavior and subject grades. Among them, student information includes student gender, age, family status and parents' education level, student behavior includes student study time, student absence number, etc., and subject grades include mathematics, Portuguese and G1, G2, G3 three semesters. There are 1044 student data records in total. The experiment uses a ten-fold cross-division data set and training set, and takes the mean as the experimental result. The data content of some data sets is shown in **Table 3**.

Table 3. Partial experimental data.

Student Number	Sex	Age	Medu	Fedu	Studytime	Absences	G1	G2	G3
1	F	18	4	4	2	4	0	11	11
2	M	16	4	3	2	6	12	12	13
3	M	15	2	2	3	0	14	14	15
4	F	16	4	4	3	10	13	13	14
5	M	16	3	1	4	2	13	11	11
6	F	16	2	2	4	1	13	13	13
7	F	16	4	4	1	4	12	13	13
8	M	15	2	1	2	4	4	9	4
9	F	19	0	1	2	0	9	10	11
10	F	18	3	2	3	10	12	11	11

Table 3 shows 9 characteristic data of 10 students, which are gender, age, parents' education level, study time, number of absences and scores in three semesters. The range of parents' education level is 0–4, where 0 represents no education, 1 represents primary school education, 2 represents middle school education, 3 represents high school education, and 4 represents college education and above. The range of study time is 1–4, where 1 represents less than two hours, 2 represents 2–2 h, 3 represents 5–10 h, and 4 represents more than 10 h. The score range of three semesters is 0–20.

3.2. Data preprocessing

3.2.1. Missing value processing

For student data, it is normal to have missing data. This paper studies the processing of missing values in three different situations. If the missing data is of numerical type, it is filled with the mean of the column. If the missing data is of categorical type, it is filled with the mode of the column. If the missing values of a row exceed 30%, the row is deleted.

Mean filling formula:

$$X_{\text{mean}} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{18}$$

In Equation (18), X_{mean} is the mean of the feature, x_i is the individual values of the feature, and n is the number of values.

Mode fill formula:

$$\text{Mode}(x) = \text{argmax}_x \text{count}(x) \tag{19}$$

In Equation (19), $\text{Mode}(x)$ is the mode of the feature, that is, the value with the highest frequency of occurrence.

3.2.2. Outlier processing

Outliers refer to extreme values that deviate from most other data in the data set, and may also be values that do not conform to the data type. This article has two ways to handle outliers. If the data type of the outlier is correct and the *Z*-Score is lower than the threshold, the mean is replaced. If the data type is wrong or the *Z*-Score is higher than the threshold, the outlier is directly deleted.

Z-Score calculation formula:

$$z = \frac{x - \mu}{\sigma} \tag{20}$$

In Equation (20), x is the data point, μ is the mean of the feature, and σ is the standard deviation.

3.2.3. Feature normalization

In order to compare different features at the same scale, this paper normalizes the data set and scales the feature values to the range of [0,1]. The normalization formula is:

$$x_{\text{norm}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \tag{21}$$

3.3. Experimental evaluation indicators

Accuracy:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{22}$$

Among them, TP (True Positive) indicates the number of correctly predicted positive classes, TN (True Negative) indicates the number of correctly predicted negative classes, FP (False Positive) indicates the number of incorrectly predicted positive classes, and FN (False Negative) indicates the number of incorrectly predicted negative classes.

Accuracy:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{23}$$

Recall:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{24}$$

$F1$ score:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{25}$$

3.4. Experimental design

The primary objective of this experiment is to establish a model that integrates factor space mathematical theory with the XGBoost algorithm to:

Accurately predict students' final grades in the upcoming semester.

Classify students into different grade categories.

Identify potential problems or difficulties in the learning process.

Conduct correlation analysis to uncover the relationships between students' grades and their features, such as demographic, behavioral, and academic attributes.

Experimental Steps:

Preprocessing the Experimental Data:

Handle missing values, outliers, and noise in the dataset.

Standardize or normalize numerical data to ensure compatibility with the machine learning model.

Encode categorical features (e.g., gender, family status) into numerical values using methods such as one-hot encoding or label encoding.

Modeling with Factor Space Mathematical Theory:

Use factor space mathematical theory to model multiple influencing factors (e.g., student background, behavior, and academic performance).

Perform feature engineering to extract key features and map them into a high-dimensional space, capturing complex and interrelated relationships between these factors.

Dimensionality Reduction with PCA:

Apply the Principal Component Analysis (PCA) method to project high-dimensional features into a low-dimensional space while preserving the most relevant information.

This step reduces redundancy and noise in the dataset, enhancing computational efficiency and mitigating overfitting risks.

Training the XGBoost Model:

Input the reduced features into the XGBoost model for training.

Use hyperparameter tuning to optimize the model parameters, such as the learning rate, number of trees, maximum tree depth, and subsample ratio.

Model Evaluation and Verification:

Evaluate the model's performance using key metrics such as:

Accuracy: The proportion of correctly classified students.

Precision, Recall, and *F*1-Score: For evaluating the balance between false positives and false negatives.

Mean Squared Error (MSE): To measure prediction errors for continuous grade values.

Use cross-validation to ensure the model's robustness and generalization ability.

Result Analysis:

Analyze the results of the model to assess its predictive accuracy and classification capabilities.

Perform feature importance analysis to identify which features have the most significant impact on student grades.

Interpret the correlation analysis to provide actionable insights for teachers and administrators to improve teaching strategies and student learning outcomes.

This step-by-step framework ensures a comprehensive approach to modeling, training, and evaluating the performance of the proposed method while addressing the challenges of complex, multidimensional educational data.

4. Result analysis

4.1. Analysis of model performance evaluation indicators

To evaluate the performance improvements or limitations of the proposed model, which integrates factor space theory with XGBoost, in predicting final exam scores, this study compares its performance metrics with those of traditional machine learning models, including XGBoost, SVM, and decision tree. The comparison focuses on key evaluation indicators such as accuracy, precision, recall, and *F*1 score. The results are visualized to highlight the differences and advantages of the proposed model over traditional approaches.

Figure 3 presents the performance evaluation metrics of four different models. The horizontal axis represents the score for each metric, while the vertical axis lists the respective metrics. Based on the data illustrated in the figure, the proposed model, which integrates factor space theory with XGBoost, outperforms the traditional XGBoost model, SVM model, and decision tree model across all four performance indicators, particularly in accuracy and *F*1 score.

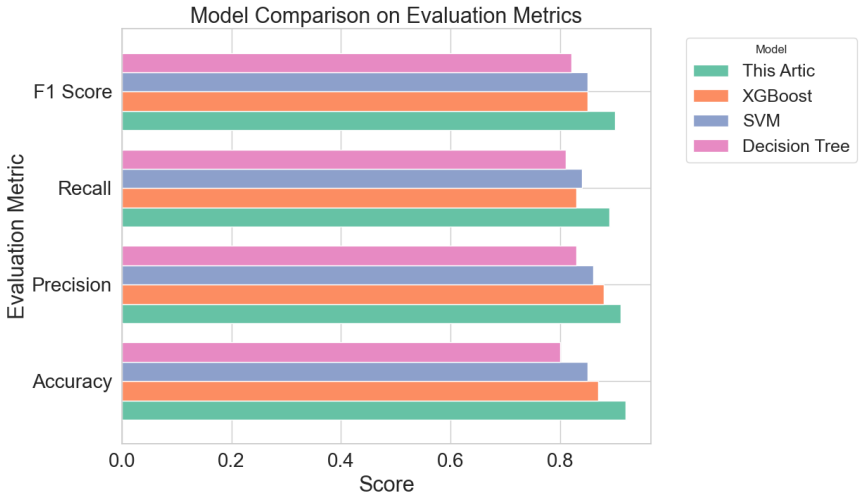

Figure 3. Model evaluation index analysis diagram.

The accuracy of the proposed model is 0.92, surpassing the traditional XGBoost model (0.87), SVM model (0.85), and decision tree model (0.80). This demonstrates the model's superior ability to correctly predict a higher proportion of samples. Additionally, the proposed model achieves an *F*1 score of 0.90, outperforming the traditional XGBoost model (0.85), decision tree model (0.82), and SVM model (0.85).

This highlights the model's ability to strike a better balance between precision and recall.

Overall, the proposed model exhibits significant advantages in performance, indicating its capability to provide more accurate and reliable predictions for higher vocational education data classification tasks.

4.2. Analysis of model generalization ability

The error performance of the test model on different data sets can observe the impact of model complexity on overfitting. In order to analyze the generalization ability of the model studied in this paper, the maximum tree depth is set to 20, and the error analysis graph is drawn by observing the different error performances of the model on the training set and the test set.

Figure 4 illustrates the error variation of the proposed model on both the training set and the test set as the model complexity increases, represented by the number of trees added to the model. The horizontal axis denotes model complexity, while the vertical axis represents the mean square error (MSE).

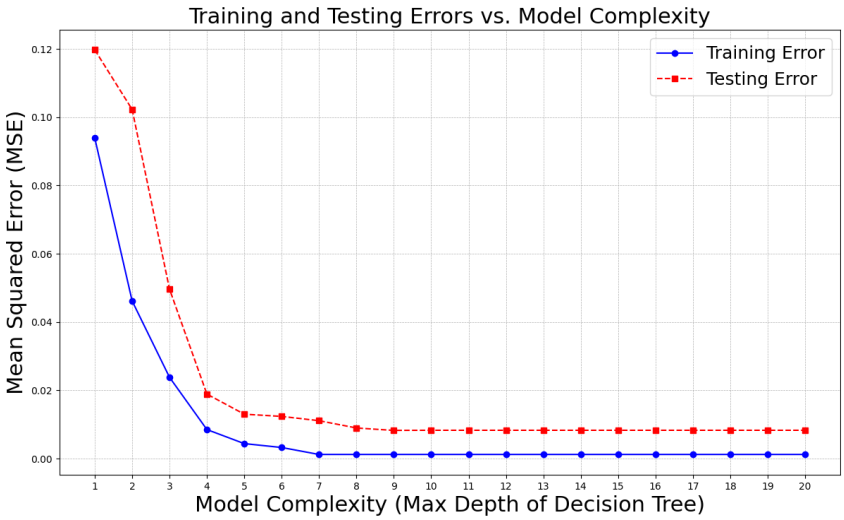

Figure 4. Error analysis.

From the figure, it can be observed that when the model complexity is low, the MSE for both the training set and the test set is relatively high. This is because a simple model lacks the capacity to capture meaningful patterns from the data, leading to underfitting. As the model complexity increases, the MSE for both sets starts to decrease, indicating that the model becomes better at fitting the data and learning relevant patterns. Interestingly, the figure shows that after the MSE for the test set reaches its minimum, there is no noticeable upward trend, which suggests that the model does not suffer from overfitting. This demonstrates that the proposed model possesses strong generalization ability, enabling it to maintain good performance when applied to new, unseen data.

4.3. Feature importance analysis

In order to study the correlation between the predicted final exam and the main features, the study visualized the prediction of the fourth semester final grade and the ranking of the main feature correlation in the model, and analyzed the importance of the main features in the student data on the final grade. The main characteristics included the duration of study, number of absences, parents' education level and grades in the previous three semesters.

Figure 5 presents an analysis of the importance of key features in predicting the final grades of the fourth semester. The horizontal axis represents the feature importance probability, while the vertical axis lists the feature names. According to the figure, the two most influential features for predicting final grades are the number of absences (importance: 0.530) and the first semester grades (importance: 0.283).

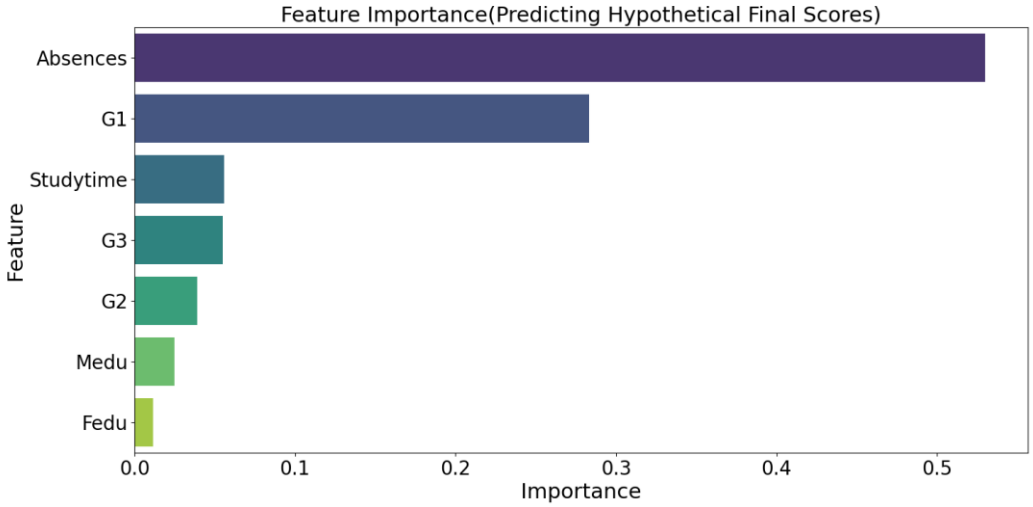

Figure 5. Feature importance analysis.

The results suggest that the number of absences has the highest importance, likely because a student's attendance rate is positively correlated with academic performance. Missing classes may lead to students not keeping up with key course content, ultimately affecting their grades. Similarly, the first semester grades rank as the second most important feature. This is because a student's early academic performance often reflects their learning abilities and habits, which tend to influence subsequent results in a cumulative learning process.

However, it is important to note that this correlation is derived from the specific dataset used in this study. In practical applications, it is necessary to have a deeper understanding of the characteristics of the dataset and the underlying influencing factors. In the context of higher vocational education, analyzing the importance of features related to students' final exam scores can provide actionable insights, allowing educators to implement targeted interventions and improve students' academic outcomes.

4.4. Analysis of the actual impact of the prediction results

In order to analyze whether the model can classify students' grades in the process of higher vocational education and predict whether the difficulties or problems that students may encounter in the learning process are solved in the students' subsequent learning, that is, whether the students' grades are improved. In addition, to analyze whether the model has an impact on the teaching quality, 20 students who have been predicted were randomly selected and a bar chart comparing their grades before and after the prediction was drawn.

Figure 6 provides a comparison of students' grades before and after addressing problems identified through the model's predictions regarding the learning process. The horizontal axis represents the student numbers (1 to 20), while the vertical axis reflects the corresponding grades.

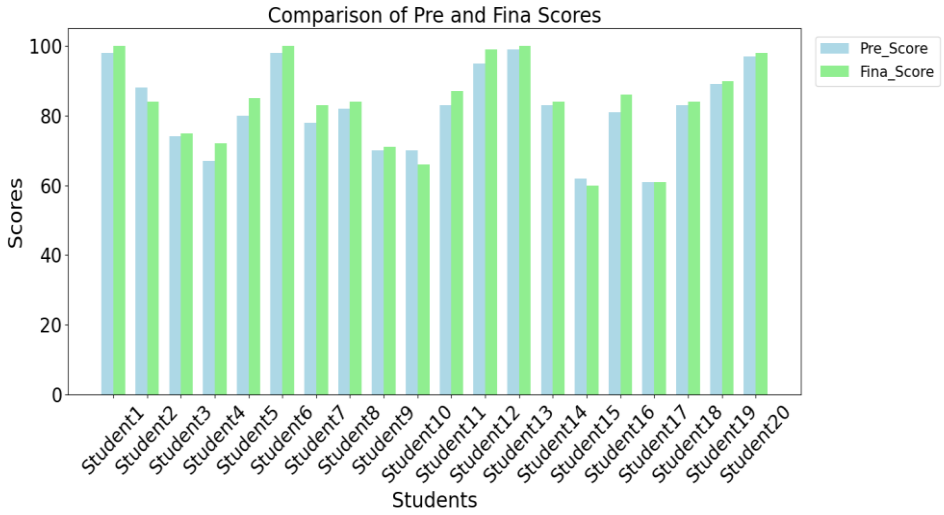

Figure 6. Comparison of results before and after.

The data in the figure reveals that among the 20 randomly selected students, 16 students experienced grade improvements ranging from 1 to 5 points, 1 student's grades remained unchanged, and 3 students experienced grade decreases of 2 to 4 points. Upon further analysis, the average grade of the students before intervention was 81.9, which increased to 83.45 after intervention, representing an average improvement of 1.55 points.

Although the improvement in average scores appears modest, it reflects a positive and meaningful change. This result indicates that addressing the predicted issues in the learning process had a generally favorable impact on the students' academic performance. It further demonstrates that the model can effectively identify and predict learning-related problems, enabling targeted interventions that contribute to improved grades in the context of higher vocational education. This underscores the practical value of the model in enhancing students' learning outcomes and supports its role in driving better academic performance.

4.5. Discussion on the integration of the model with biomechanics studies

The model presented in this study, which combines factor space mathematical theory with the XGBoost algorithm, offers a robust framework for analyzing complex multidimensional data. This methodology can also be extended to biomechanical studies, where data complexity and feature interactions are key challenges. For example, in biomechanical research, various factors such as anatomical measurements, movement patterns, muscle activity, and external environmental influences often interact in nonlinear ways. By leveraging factor space theory, these multidimensional factors can be mapped into a high-dimensional space to capture complex relationships, while the XGBoost algorithm can provide accurate predictions and classifications of biomechanical behaviors, such as injury risk, rehabilitation progress, or performance optimization. This integration could significantly enhance the precision and generalization capabilities of biomechanical models, especially in areas like sports science and physical therapy.

Moreover, the feature engineering and dimensionality reduction techniques used in this study, such as PCA, can be applied to filter out noise and extract critical biomechanical variables. For instance, in predicting movement patterns or diagnosing musculoskeletal issues, high-dimensional data such as kinematic and kinetic measurements could be reduced to their most relevant components, enabling efficient computation and reducing overfitting. The ability of this model to predict problems or classify behaviors with high accuracy (as demonstrated in educational data analysis) suggests its potential to identify key biomechanical factors influencing outcomes like joint stability or muscle fatigue. This cross-disciplinary application not only highlights the adaptability of the model but also provides a foundation for developing advanced tools for personalized biomechanics, bridging the gap between data-driven predictions and real-world physiological insights.

5. Conclusion

This study introduces a learning model based on factor space mathematical theory and XGBoost for application in higher vocational education. Using the public data set Student Performance Data, classification and prediction experiments were conducted to evaluate the model's performance. The results demonstrate that, compared to the standalone XGBoost model and other traditional machine learning models, the proposed model exhibits superior classification accuracy, stronger prediction capability, and robust generalization across different datasets, making it well-suited to diverse educational data scenarios. Furthermore, the integration of factor space theory allows the model to effectively handle multidimensional and complex interactions, addressing key limitations of traditional approaches.

However, while the model achieved high accuracy and an excellent $F1$ score in this study, the experimental validation was limited to a single dataset. Although the combination of factor space theory and the XGBoost algorithm demonstrated strong potential, the model's optimization and tuning were constrained. Future research should focus on conducting extensive experiments across a wider range of datasets to validate the model's stability and applicability in diverse educational contexts. Additionally, further optimization and refinement of the model will be explored to

enhance its performance, with the ultimate goal of providing more accurate, scalable, and practical solutions for educational data analysis and prediction.

Author contributions: Conceptualization, WL and GL; methodology, WL; software, WL; validation, WL, GL and ZW; formal analysis, WL; investigation, WL; resources, WL; data curation, WL; writing—original draft preparation, WL; writing—review and editing, GL and ZS; visualization, WL; supervision, ZW; project administration, GL and ZS; funding acquisition, ZW. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Education Department of Sichuan Province (SCJG23A259).

Ethical approval: Not applicable.

Conflict of interest: The authors declare no conflict of interest.

References

1. Mason G. Higher education, initial vocational education and training and continuing education and training: Where should the balance lie. Journal of Education and Work. 2020; 33(7–8): 468–490.
2. Hariri RH, Fredericks EM, Bowers KM. Uncertainty in big data analytics: Survey, opportunities, and challenges. Journal of Big Data. 2019; 6(1): 1–16.
3. Holmes W, Tuomi I. State of the art and practice in AI in education. European Journal of Education. 2022; 57(4): 542–570.
4. Luan H, Tsai CC. A review of using machine learning approaches for precision education. Educational Technology & Society. 2021; 24(1): 250–266.
5. Marques LS, Gresse von Wangenheim C, Hauck JCR. Teaching machine learning in school: A systematic mapping of the state of the art. Informatics in Education. 2020; 19(2): 283–321.
6. James CA, Wheelock KM, Woolliscroft JO. Machine learning: The next paradigm shift in medical education. Academic Medicine. 2021; 96(7): 954–957.
7. Berens J, Schneider K, Gortz S, et al. Early Detection of Students at Risk--Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods. Journal of Educational Data Mining. 2019; 11(3): 1–41.
8. Ikawati Y, Al Rasyid MUH, Winarno I. Student behavior analysis to predict learning styles based felder silverman model using ensemble tree method. EMITTER International Journal of Engineering Technology. 2021; 9(1): 92–106.
9. Ouatik F, Erritali M, Ouatik F, Jourhmane M. Predicting student success using big data and machine learning algorithms. International Journal of Emerging Technologies in Learning (iJET). 2022; 17(12): 236–251.
10. Wu TK, Huang SC, Meng YR. Evaluation of ANN and SVM classifiers as predictors to the diagnosis of students with learning disabilities. Expert Systems with Applications. 2008; 34(3): 1846–1856.
11. Amin S, Uddin MI, Mashwani WK, et al. Developing a personalized E-learning and MOOC recommender system in IoT-enabled smart education. IEEE Access. 2023; 11: 136437–136455.
12. Alshurafat H, Al Shbail MO, Masadeh WM, et al. Factors affecting online accounting education during the COVID-19 pandemic: An integrated perspective of social capital theory, the theory of reasoned action and the technology acceptance model. Education and Information Technologies. 2021; 26(6): 6995–7013.
13. Delen D, Topuz K, Eryarsoy E. Development of a Bayesian Belief Network-based DSS for predicting and understanding freshmen student attrition. European journal of operational research. 2020; 281(3): 575–587.
14. Li R, Du C, Du W, et al. Research on comprehensive evaluation model of physical education teaching quality based on multivariate data. Journal of Sport Psychology. 2022; 31(1): 235–244.
15. Godwin A, Benedict B, Rohde J, et al. New epistemological perspectives on quantitative methods: An example using topological data analysis. Studies in Engineering Education. 2021; 2(1): 16–34.

16. Mubarak AA, Cao H, Hezam IM, Hao F. Modeling students' performance using graph convolutional networks. Complex & Intelligent Systems. 2022; 8(3): 2183–2201.
17. Yakubu MN, Dasuki SI. Factors affecting the adoption of e-learning technologies among higher education students in Nigeria: A structural equation modelling approach. Information Development. 2019; 35(3): 492–502.
18. Osman AIA, Ahmed AN, Chow MF, et al. Extreme gradient boosting (XGBoost) model to predict the groundwater levels in Selangor Malaysia. Ain Shams Engineering Journal. 2021; 12(2): 1545–1556.
19. Kavzoglu T, Teke A. Predictive Performances of ensemble machine learning algorithms in landslide susceptibility mapping using random forest, extreme gradient boosting (XGBoost) and natural gradient boosting (NGBoost). Arabian Journal for Science and Engineering. 2022; 47(6): 7367–7385.
20. Arabameri E. The Evolution of Motor Behavior: Lessons from Past Research and Future Prospects. Health Nexus. 2024; 2(4): 134–151.
21. Padilla BO. Deep state-space modeling for explainable representation, analysis, and forecasting of professional human body dynamics in dexterity understanding and computational ergonomics [Doctoral dissertation]. Université Paris sciences et lettres; 2023.
22. Ebers MR. Machine learning for dynamical models of human movement [Doctoral dissertation]. University of Washington; 2023.
23. Donmazov S, Saruhan EN, Pekkan K, Piskin S. Review of machine learning techniques in soft tissue biomechanics and biomaterials. Cardiovascular Engineering and Technology. 2024; 15: 1–28.
24. Mishra N, Habal BGM, Garcia PS, Garcia MB. Harnessing an AI-Driven Analytics Model to Optimize Training and Treatment in Physical Education for Sports Injury Prevention. In: Proceedings of the 2024 8th International Conference on Education and Multimedia Technology; 22–24 June 2024; Tokyo, Japan. pp. 309–315.
25. Mishra PK, Fasshauer GE, Sen MK, Ling L. A stabilized radial basis-finite difference (RBF-FD) method with hybrid kernels. Computers & Mathematics with Applications. 2019; 77(9): 2354–2368.
26. Gewers FL, Ferreira GR, De Arruda HF, et al. Principal component analysis: A natural approach to data exploration. ACM Computing Surveys (CSUR). 2021; 54(4): 1–34.
27. Hasan BMS, Abdulazeez AM. A review of principal component analysis algorithm for dimensionality reduction. Journal of Soft Computing and Data Mining. 2021; 2(1): 20–30.
28. Engle RF, Ledoit O, Wolf M. Large dynamic covariance matrices. Journal of Business & Economic Statistics. 2019; 37(2): 363–375.
29. Zhang Z, Jung C. GBDT-MO: Gradient-boosted decision trees for multiple outputs. IEEE transactions on neural networks and learning systems. 2021; 32(7): 3156–3167.
30. Mistry M, Letsios D, Krennrich G, et al. Mixed-integer convex nonlinear optimization with gradient-boosted trees embedded. INFORMS Journal on Computing. 2020; 33(3): 1103–1119.
31. Moradi R, Berangi R, Minaei M. A survey of regularization strategies for deep models. Artificial Intelligence Review. 2020; 53(6): 3947–3986.
32. Bejani MM, Ghatee M. A systematic review on overfitting control in shallow and deep neural networks. Artificial Intelligence Review. 2021; 54(8): 6391–6438.
33. Chen J, Pu Y, Guo L, et al. Second-order optimization methods for time-delay autoregressive exogenous models: Nature gradient descent method and its two modified methods. International Journal of Adaptive Control and Signal Processing. 2023; 37(1): 211–223.
34. Guo J, Fu H, Pan B, Kang R. Recent progress of residual stress measurement methods: A review. Chinese Journal of Aeronautics. 2021; 34(2): 54–78.